“Users of scooters must transfer to a fixed seat”

(Notice in Brisbane Taxi)

Introduction to BUGS in Genetic Epidemiology

“**B**ayesian Inference **U**sing **G**ibbs **S**ampling”

Lindon Eaves, Tim York

Leuven, August 2008.

Government Health Warning

Potential users are reminded to be extremely careful if using this program for serious statistical analysis. We have tested the program on quite a wide set of examples, but be particularly careful with types of model that are currently not featured. If there is a problem, *WinBUGS* might just crash, which is not very good, but it might well carry on and produce answers that are wrong, which is even worse. Please let us know of any successes or failures.

Beware: MCMC sampling can be dangerous!

From: WinBUGS manual Page 1

Goals

- Give overview of MCMC – what is it and why use it.
- Give some simple examples of building blocks of applications: means, variances, twin correlations.
- Show how building blocks can be developed for more complex (“interesting”, “challenging”) problems that give a flavor of what MCMC might be used for.

Apology

I don’t know much and probably don’t know what I am talking about. But I hope others will see possibilities and help deepen understanding.

Thanks

Allattin Erkanli

Nick Martin

Staff and students QIMR

Critical Source: MRC BUGS Project

http://www.mrc-bsu.cam.ac.uk/bugs/

Why bother?

- Intellectual: Different (“Bayesian”) way of thinking about statistics and statistical modeling.
- Pragmatic:“Model Liberation” – can do a bunch of cool new stuff much more easily than within ML framework.
- Pragmatic: Learn more about data for less computational effort
- Pragmatic: “Fast”

Payoff

- Estimate parameters of complex models
- Obtain subject parameters (e.g. “genetic and environmental factor scores”) at no extra-cost
- Obtain confidence intervals and other summary statistics (s.e’s, quantiles etc) at no extra cost.
- Automatic imputation of missing data (“data augmentation”)
- Fast (35 item, IRT in 500 twins with covariates takes about 1-2 hours on laptop).
- Insight, flexibility

Generally:

Seems to help with models that require multi-dimensional integration to compute likelihood.

Some applications

- Non-linear latent variables (GxE interaction).
- Multidimensional, multi-category, multi-item IRT in twins.
- Genetic effects on developmental change in multiple indicators of puberty (latent growth curves).
- Hierarchical mixed models for fixed and random effects of G, E and GxE in multi-symptom (“IRT”) twin data.
- Genetic survival models
- Mixture models.
- Estimating (genetic?) factor scores – (?genetic) counseling, c.f.. estimation of breeding values
- Imputation of missing values
- Model selection (“Reversible Jump” MCMC)
- Non-normal outcomes (e.g. symptom counts)
- Awkward designs: incomplete randomization, batches

This introduction

- Introduce ideas
- Practice use of WinBUGS
- Run some basic examples
- Look at application to genetic IRT
- Other stuff?

Some resources:

Gilks WR, Richardson S, Spiegelhalter DJ (1996)
Markov Chain Monte Carlo in Practice. Boca Raton, Chapman & Hall,

Gelman A, Carlin JB, Stern HS, Rubin DB. (2004)
Bayesian Data Analysis (2nd Ed,) Boca Raton, Chapman & Hall.

Spiegelhalter DJ, Thomas A, Best N, Lunn D. (2004). *WinBUGS User Manual Version 1.4.1.* Cambridge, England. MRC BUGS project. [Downloaded with WinBUGS – also Examples Vols. I and II]

Maris, G and Bechger, T.M. (2005). An Introduction to the DA-T Gibbs Sampler for the Two-Parameter Logistic (2PL) Model and Beyond. *Psicol´ogica: **26**, 327-352.*
http://www.uv.es/~revispsi/articulos2.05/8-MARIS.pdf

http://www.helsinki.fi/~mjlaine/id2000ml.pdf Applications of reversible jump MCMC. Marko Laine. University of Helsinki Department of Mathematics.

Lunn, D.J.; Whittaker, J.C.; Best, N.;
A Bayesian toolkit for genetic association studies.
Genet Epidemiol, 2006; 30(3):231-47

Basic Ideas

- Bayesian Estimation (vs. ML)
- “Monte Carlo”
- “Markov Chain”
- Gibbs sampler

(Maximum) Likelihood

- Compute (log-) likelihood of getting data given values of model parameters and assumed distribution
- Search for parameter values that maximize likelihood (“ML” estimates)
- Compare models by comparing likelihoods
- Obtain confidence intervals by contour plots (i.e. repeated ML conditional on selected parameters)
- Obtain s.e.’s by differentiating L

Problem with ML

- Many models require integration over values of latent variables (e.g. non-linear random effects)

- Integrate to evaluate each likelihood and derivatives for each parameter

- “Expensive” when number of dimensions is large (?days), especially for confidence intervals.

Maximum Likelihood (ML) “Thinks” (*theoretically*) about parameters and data separately: P(data|parameters)

“Thinks” (*practically*) of integration, searching and finding confidence intervals as separate numerical problems (quadrature, e.g. Newton-Raphson, numerical differentiation).

Markov Chain Monte Carlo (MCMC, MC^2) “Thinks” (*theoretically)* that there is no difference between parameters and data – seeks *distribution* of parameters given data – P(parameters|data) {*Bayesian* estimation}

“Thinks” *(practically)* that integration, search and interval estimation constitute a single process addressed by a single unifying algorithm {*Gibbs Sampling}*

“Parameter”

Anything that isn’t data: means, components of variance, correlations. But also subjects’ scores (not just distributions) on latent traits (genetic liabilities, factor scores), missing data points.

Basic approach

- “Bayesian” = Considers joint distribution of parameters and data
- “Monte Carlo” = Simulation
- “Markov Chain” = Sequence of simulations (“time series”) designed to converge on samples from posterior distribution of θ given *D*
- “Gibbs sampler” = method of conducting simulations – cycles through all parameters simulating new value of parameter conditional on *D* and every other parameter

“Bayesian”

- Considers joint distribution of all parameters (θ) and data (*D*): P(θ.D)
- Seeks “posterior distribution” of θ given *D:*

  P(θ|*D*)
- Need to know “prior” distribution P(θ), but don’t.
- Start out by assuming some prior distribution (“uniformative” priors – i.e. encompassing wide range of possible parameter values) and seek to refine using data.

“Monte Carlo”

- Computer simulation of unknown parameters (“nodes”) from assumed distribution (“computer intensive”).
- If distribution is assumed [e.g. mean and variance] then successive simulations represent samples from the assumed distribution i.e. Can estimate “true” distribution from large number of (simulated) samples – can get any properties of distribution (means, s.d.s, quantiles) to any desired degree of precision.

“Markov Chain”

- “True” prior distribution P(θ) unknown.
- “Markov Chain” – series of outcomes (e.g. sets of data points) each contingent on previous outcome – under certain conditions reach a sequence where underlying distribution does not change (“stationary distribution”).
- Start with assumed prior distribution and construct (simulate) Markov Chain for given *D* that converges to samples from posterior distribution: P(θ|D).
- Then use (large enough) set of samples from stationary distribution to characterize properties of desired posterior distribution.

“Gibbs Sampler”

- Algorithm for generating Markov chains from multiple parameters conditional on *D*.
- Takes each parameter in turn and generates new value conditional on the data and every other parameter. Cycle through all parameters (“one iteration”) and repeat until converge to stationary distribution.

Idea of Gibbs Sampler

Suppose want to draw samples (x,y) from bivariate normal:

Could do it by sampling x from univariate normal and then sample y from univariate normal of y|x, then new x|y….

Sucessive samples of (x,y) are from bivariate normal

WinBUGS

- Free
- Simple language (“R-like”)
- [“Open” Version: “Open BUGS”]
- PC version (BUGS also available for mainframe
- Graphical interface (OK but usually easier to write or adapt existing code)
- Well documented, good examples

Problems

- Convergence criteria (“mixing”)
- Model comparison
- ? Sensitivity to priors
- Model identification
- Error messages sometimes obscure
- Data set-up can be a pain
- Can have problems (Latent class analysis)

Bottom line

If you can figure how you would simulate it, you can probably “BUGS-it.”

Need to be clear and explicit about model and assumptions.

Today

- Tour BUGS
- Run some simple examples of easy problems
- Illustrate application to IRT

"Ladies and gentlemen…
start your engines…"